# Computación de Alto Desempeño y Datos Masivos

Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, Cristian Tissera, Marcelo Alaniz, Ruben Apolloni, Alicia Castro, Fabiana Piccoli

LIDIC- Universidad Nacional de San Luís
San Luís, Argentina
{mdbarrio, omlopres, ncmiran, ptissera, rubenga, adcastro, mpiccoli}@unsl.edu.ar

## Resumen

*Este trabajo presenta las líneas de trabajo orientadas al desarrollo y uso de técnicas de Computación de Alto Desempeño en problemas con datos masivos o problemas Big-Data. Por un lado se aborda el estudio de arquitecturas, el desarrollo de metodologias y el diseño de herramientas de computación de alto desempeño para dar solución a problemas que deben tratar con gran cantidad de datos en forma eficiente. Adicionalmente, en estos contextos es crucial la predicción de la performance , los que nos conduce a la modelización de algoritmos y a la estimación de sus costos sobre arquitecturas de alta performance.*

**Palabras clave:** Datos Masivos, Computación de Alto Desempeño, Arquitecturas Multicore y Many core.

## Contexto

Esta propuesta de trabajo se lleva a cabo dentro del proyecto de investigación “Tecnologías Avanzadas aplicadas al Procesamiento de Datos Masivos” y de otros dos proyectos asociados: a) Proyecto binacional CAPG-BA 66/13 entre la Universidad Nacional de San Luis (UNSL) y la Universidad Federal de Pernambuco (UFPE), Recife, Brasil y b) Programa Regional STIC-AMSUD entre MINCYT (Argentina), Inria (Francia) y ANII (Uruguay).

El proyecto de investigación se desarrolla en el marco del Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC), de la Facultad de Ciencias Físico, Matemáticas y Naturales de la UNSL.

## Introducción

Con el uso masivo de internet, estamos en presencia de un fenómeno donde la rápida aceleración tanto del crecimiento del volumen de datos capturados y almacenados, como la creciente variación en los tipos de datos requeridos, hace que las técnicas tradicionales para el procesamiento, análisis y obtención de información útil deban ser redefinidas para formular nuevas metodologías de abordaje.

Trabajar con grandes volúmenes de datos (llamados datos masivos a gran escala) implica un gran desafío debido a

la necesidad de explorar un universo de nuevas tecnologías, las cuales no sólo hacen posible la obtención y procesamiento de los datos sino también realizan su gestión en un tiempo razonable. Este crecimiento es algo cotidiano y obedece a la proliferación de páginas web, aplicaciones de imagen y vídeo, redes sociales, dispositivos móviles, sensores, internet de las cosas, etc.. Todas ellas capaces de generar según IBM, más de 2.5 quintillones de bytes diariamente. Ejemplos cotidianos de datos masivos son el número de imágenes subidas diariamente a las redes sociales (300 millones en Facebook, 45 millones en Instagram), los videos vistos por día en YouTube (2 billones), la cantidad de mensajes de texto enviados por un adolescente en un mes (4762), la cantidad mensual de búsquedas en Twitter, el tráfico mundial en internet, entre otros. Esto no sólo es aplicable a las actividades desarrolladas diariamente en internet, sino también en aquellas relacionadas a fenómenos naturales como el clima o datos sismográficos, entornos referidos a la salud, la seguridad o, por supuesto, al ámbito empresarial.

Para trabajar con estos datos a gran escala, es necesario tener en cuenta que la obtención de información útil será a partir de datos no estructurados como texto, audio, imagen y video. Es por ello que se debe considerar la aplicación de nuevas metodologías para el procesamiento eficiente y eficaz de estos grandes volúmenes de datos. Además, como cada uno de los procesos involucrados en el proceso de obtener información a partir de datos masivos implica un gran número de problemas computacionalmente costosos, el uso de nuevas técnicas y arquitecturas puede contribuir a mejorar su rendimiento; es por ello que la búsqueda y selección de técnicas de computación de altas prestaciones (HPC) en cada etapa o proceso involucrado permitirá resolver con eficiencia cada uno de los objetivos a planterse.

La preocupación por tratar datos a gran escala llevó a crear algoritmos y modelos de programación distribuidos y paralelos, como MapReduce[DG04], Hive[CWR12] e Impala[R13] que permiten procesar terabytes de información sin necesidad de cambiar las estructuras de datos subyacentes. Si los datos analizados se distribuyen entre varios servidores en Internet, entonces las consultas de búsqueda deberán dirigirse a estos servidores en paralelo. En particular el framework de Google MapReduce es un modelo de programación diseñado para procesar conjuntos de datos a gran escala en una modalidad orientada al lote (batch) o procesamiento online. Tecnologías como Hadoop [W09] adoptaron estos modelos de programación, permitiendo el procesamiento distribuido y paralelo de grandes cantidades de datos a través de clusters de servidores de bajo costo los cuales almacenan, procesan y transforman los datos. Una característica importante de Hadoop es la partición de datos y de cómputo a través de muchos (miles) de hosts, y la ejecución de las aplicaciones en

paralelo muy cerca de sus datos. Hadoop ofrece escalabilidad y flexibilidad para almacenar grandes volúmenes de datos heterogéneos sin necesidad de conocer a priori cómo se los va a procesar [W09].

La reducción significativa de los tiempos de procesamiento ha conducido a mayores expectativas, dando como resultado, por ejemplo, el enfoque conocido como Big Data [MC13,MCJ13]. Esta tecnología no sólo es un proceso para almacenar y recuperar rápidamente petabytes o exabytes de datos desde una data warehouse sino también involucra procesos con la capacidad de tomar mejores decisiones, reduciendo el tiempo entre la ocurrencia de evento en algún lugar del mundo y la capacidad de reaccionar a ese evento [L13]. Se trata de la combinación y el análisis de datos para que un proceso o persona pueda tomar la acción correcta, en el momento adecuado y en el lugar correcto. Big data tiene tres dimensiones volumen, variedad y velocidad y dentro de cada una de estas tres dimensiones se presenta una amplia gama de aspectos [N13, B13] los que serán contemplados en el presente proyecto.

La presente propuesta tiene como objetivo aplicar técnicas HPC en las etapas del proceso de obtención de información a partir de datos masivo considerando arquitecturas multi y many core como arquitecturas subyacentes.

## Líneas de Investigación, Desarrollo e Innovación

En la actualidad, la computación de alto desempeño, más concretamente computación paralela, está siendo utilizada en multitud de campos para el desarrollo de aplicaciones y el estudio de problemas que requieren gran capacidad de cómputo, ya sea por el gran tamaño de los problemas que abordan o por la necesidad de disminuir el tiempo de respuesta. Para obtener un sistema de alta performance es necesario una infraestructura conformada por paradigmas, metodologías y tecnologías, aspectos que constituyen líneas abiertas en el ámbito de la investigación científica y tecnológica.

Esta línea de investigación tiene como objetivo realizar un análisis de la aplicación de modelos y técnicas HPC en datos masivos no estructurados. Para lograrlo nos planteamos tres líneas de investigación, ellas son:

- Modelos y paradigmas de computación de alto desempeño: la programación paralela involucra muchos aspectos, los cuales no se presentan en la programación convencional. El diseño de un sistema paralelo tiene que considerar entre otras cosas, el tipo de arquitectura sobre la cual se va a ejecutar el programa, las necesidades de tiempo y espacio requeridas por la aplicación; las técnicas y estructuras

de programación paralela adecuadas para implementarla; y la forma de coordinar y comunicar las diferentes unidades computacionales dedicadas a resolver conjuntamente el problema. Además, en la última década las arquitecturas paralelas han evolucionado drásticamente (clusters de pc, procesadores multicore, procesadores manycore) existiendo un desafío adicional para las aplicaciones paralelas que consiste en explotar tales arquitecturas a su máximo potencial.

- Algoritmos y Estrategias de alta performance en grandes volúmenes de datos: un punto a tener en cuenta cuando lo que se desea es rendimiento, es el origen de los datos de un problema. Los enfoques tradicionales, trabajan sobre datos centralizados o sobre fragmentos del conjunto original. Con el auge de Internet esto ha cambiado, los datos son generados automáticamente en forma distribuida, como es el caso de redes de sensores "wireless", observaciones meteorológicas y astronómicas, centros de observaciones satelitales distribuidas sobre todo un país, etc. Otro inconveniente es el relacionado a la heterogeneidad de modos/formatos en la que se encuentra disponible la información y su administración eficiente. En estas dos líneas, las investigaciones tienen en cuenta la portabilidad de los desarrollos a pesar de las características propias de cada uno de los datos no estructurados.

- Simulación de Sistemas: a lo largo de esta línea se pretende investigar, especificar e implementar un modelo de simulación de alta performance. En el desarrollo de la investigación de esta línea se han desarrollado una serie de trabajos que concentran estrategias de computación paralela con técnicas de inteligencia artificial y arquitecturas de computadoras.

## Resultados y Objetivos

Como objetvos de las líneas de investigación nos planteamos facilitar el desarrollo de soluciones paralelas portables, de costo predecible y bajo consumo, capaces de explotar las ventajas de modernos ambientes de HPC a través de herramientas y "frameworks de computación" de alto nivel. Para ello será necesario proponer nuevas metodologías a ser aplicadas en cada una de las fases del tratamiento de datos masivos.

## Formación de Recursos Humanos

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento el desarrollo de 6 tesis doctorales y 4 tesis de maestría. Además se están ejecutando varias tesinas de grado.

## Referencias


[B13] M. Barlow: Real-Time Big Data Analytics: Emerging Architecture. Kindle Edition. O'Reilly Media Inc. 2013.

[CWR12] E. Capriolo, D. Wampler, J. Rutherglen: Programming Hive: Data Warehouse and Query Language for Hadoop. O'Reilly Media. 2012.

[DG04] J. Dean and S. Ghemawat: MapReduce: Simplified Data Processing on Large Clusters. Proc. Sixth Symposium on Operating System Design and Implementation, 2004.

[HP08] J. L. Hennesy and D. A. Patterson: Computer Organization & Design - The Hardware/Software Interface. Morgan Kaufmann, 4th edition, 2008.

[HW11] G. Hager, G. Wellein: Introduction to High Performance Computing for Scientists and Engineers. Chapman & Hall/CRC Computational Science. 2011.

[L13] D. Loshin: Big Data Analytics. From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph. Kaufmann . Elsevier . 2013.

[MC13] V. Mayer-Schönberger, K. Cukier: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt. 2013.

[MCJ13] V. Mayer-Schönberger, K. Cukier. A.I. Jurado: Big data: La revolución de los datos masivos. Turner. 2013.

[N13] J. Needham: Disruptive Possibilities: How Big Data Changes Everything. Kindle Edition. O'Reilly Media Inc. 2013.

[OPS01] S. Orlando, R. Perego, and F. Silvestri: Design of a Parallel and Distributed WEB Search Engine. In Proceedings of Parallel Computing (ParCo) 2001 conference. Imperial College Press, September 2001.

[R13] J.Rusell: Cloudera Impala. O'Reilly Media, Inc.2013.

[RR11] T. Rauber, G. Runger: Parallel Programming for multicore and Cluster Systems. Springer. 2011.

[V11] Valiant L.G.: A bridging model for multi-core computing. J. Comput. Syst. SCI 77(1): 154-166 (2011).

[W09] T. White, D. Cutting: Hadoop-The Definitive Guide. 2009 by O'Reilly Media 2009.

[BOM10] Broquedis, F., Clet-Ortega, J., Moreaud, S., Furmento, N., Goglin, B., Mercier, G., Thibault, S., Namyst, R.: Hwloc: A generic framework for managing hardware affinities in HPC applications. In: 18 th Euromicro Conference on Parallel, Distributed and Network-based Processing, pp. 180–186. (2010)